Panorama: A robust pangenome-based method for predicting and comparing biological systems across species

Jérôme Arnoux[1,*], Jean Mainguy[1], Laura Bry[1], Quentin Fernandez de Grado[1], Vallenet David[1], Alexandra Calteau[1]

[1] LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS

*jarnoux@genoscope.cns.fr

Abstract

Over the last decade, the expansion in the number of prokaryotic genomes available has profoundly transformed the study of genetic diversity, evolution and ecological adaptation. However, traditional approaches based on the analysis of individual genomes are showing their limitations when faced with the sheer volume of data. To overcome these limitations, the concept of the pangenome has emerged, offering an overview of genetic diversity and evolutionary dynamics within a species. In this study, we present PANORAMA, an innovative pangenomic tool designed to exploit pangenome graphs and enable interspecific comparisons to explore genomic diversity. Based on the PPanGGOLiN software suite, PANORAMA incorporates advanced methods for annotating macromolecular systems at the pangenome scale and for comparative analysis of spots of insertion between different pangenomes. We illustrate the application of PANORAMA to a Pseudomonas aeruginosa dataset, evaluating its performance against reference tools such as PADLOC and DefenseFinder. The analysis was then extended to a wider set including four Enterobacteriaceae species, demonstrating PANORAMA's ability to annotate, compare and explore the diversity and distribution of antiphage defence systems beyond the species level. This work provides a new resource for the comparative study of bacterial genomes and highlights the relevance of genome-wide approaches for deciphering the evolutionary dynamics and ecological significance of bacterial defense repertoires.

Keywords: Pangenome, Comparative genomics, anti-phage defense systems, Comparative analysis, Pseudomonas aeruginosa, Enterobacteriaceae, Bioinformatics.

Introduction

The rapid expansion of bacterial genome sequencing over the past decade has provided unprecedented opportunities to study the genetic diversity, evolution, and ecological adaptation of microbial species [1, 2]. For a significant number of species, sequences of hundreds or even thousands of strains are now available. While this wealth of information offers immense potential for discovery, it also presents significant challenges, as traditional genome-centric approaches, which focus on individual genomes, are becoming increasingly inadequate for managing and interpreting such large-scale datasets. To address these limitations, the concept of pangenome has emerged as a powerful tool. It encompasses the entire gene repertoire of a species, including core genes present in all strains and accessory genes found in only a subset, and provides a holistic view of genetic diversity and evolution within a species [3]. Pangenomics has significantly transformed microbial genomics by providing a comprehensive framework for understanding genetic diversity and functional capabilities across microbial species [4]. This approach allows researchers to investigate not only the genome of a single strain but the complete gene repertoire within a species or group of strains, the pangenome, thereby enhancing insights into microbial evolution and adaptation.

Owing to the small size of their genomes and the large number of sequences available, particularly for species of clinical interest, pangenomic analysis of microbial genomes has benefited from the early development of tools, facilitating pangenome analysis, offering visualization, comparison, and partitioning of genomic data[5]. Among these tools, **PPanGGOLiN** stands out for its unique approach to analyze pangenomes by partitioning them with a statistical algorithm [6]. PPanGGOLiN represents genomic data as a pangenome graph at the gene family level, with nodes representing homologous gene families and edges capturing their genetic contiguity, enabling the compression of information from thousands of genomes while preserving the chromosomal organization of genes. A statistical model is applied to partition gene families in persistent genome (i.e. gene families found in nearly all genomes) and variable genome, which includes the shell and cloud components corresponding to intermediate- and low-frequency gene families, respectively. PPanGGOLiN includes additional methods for the identification of Regions of Genomic Plasticity (RGPs) and their spot of insertion (panRGP method) [7] and their fine description in conserved modules (panModule method) [8], which have demonstrated their utility in identifying genomic islands and provide helpful insights into the genomic adaptability and evolution of bacteria. Despite these advances, a significant challenge remains: detecting and comparing complex macromolecular systems at the pangenome scale. In microbial genomes, the genes responsible for macromolecular systems are usually arranged in a highly structured manner, typically clustered into one or several operons composed of functionally related genes. These clusters encode coordinated systems that play essential roles in microbial life. Among them are **secretion systems**, which allow the transport of proteins and other molecules across membranes to interact with the environment or host organisms; **defense systems**, which protect the cell from foreign genetic elements; and **metabolic pathways**, which organize enzymatic reactions to efficiently produce, transform, or degrade biological molecules. Understanding the organization and diversity of these systems is key to decoding the functional capabilities of microbial genomes. Several tools have been developed to detect macromolecular systems at the genome scale, including MacSyFinder [9], PADLOC [10], and Defense Finder [11], the latter two being specialized in identifying **bacterial anti-phage immune systems**. These tools are highly effective when applied to individual genomes; however, they are not designed to detect complex systems at the pangenome level, nor to enable systematic comparisons across large genomic datasets. Tools capable of building, comparing, and functionally annotating pangenome at the scale of thousands of genomes across multiple species remain limited. Some pangenomic tools allow the functional annotation of the pangenome by incorporating results from annotation tools, as do PanGGOLiN [6], Panaroo[12], or PanTOOLs[13], or also by aligning the pangenome to a sequence database as do PPanGGOLiN[6] or PanGraph[14]. None of these tools allows searching directly into the pangenome, but only to annotate with already known results. To date, no tools are available to construct, compare, and functionally annotate pangenomes at the scale of thousands of genomes from multiple species.

Here, we introduce **PANORAMA**, a powerful computational tool designed to harness bacterial

pangenome graphs from large genomic datasets and enable comparisons across species to explore genomic diversity. Built on the PPanGGOLiN software suite, PANORAMA incorporates advanced methods for reconstructing and analyzing pangenome graphs. It offers several key features, including the ability to compare genomic contexts between pangenomes and annotate macromolecular systems at the pangenome scale. Functional annotation of biological systems is performed directly on the graph structures using rule-based models, making it possible to map and analyze complex genomic features without relying on linear genome representations. To illustrate the versatility of our approach, we focused on the comparative analysis and annotation of bacterial defense systems. Bacteria have evolved a remarkably diverse array of defense mechanisms against phages and other mobile genetic elements. These range from well-characterized systems, such as restriction-modification (R-M) systems and CRISPR-Cas complexes [15], to more recently discovered and less understood systems like BREX [16], DISARM [17], and retron-based defense systems [18]. To date, over 150 systems have been described, unveiling an unsuspected diversity of molecular mechanisms [19]. This diversity is not only taxonomically widespread but also highly dynamic; defense systems are often associated with mobile genetic elements or genomic islands and can vary extensively in composition, organization, and presence both within and between closely related species [11, 20]. Despite this complexity, large-scale comparative studies of bacterial defense systems are still scarce. Pangenome-level analyses hold great promise for revealing patterns of co-occurrence, horizontal gene transfer, evolutionary innovation, and defense strategies that are specific to particular species or lineages. In this study, we present the methodology behind PANORAMA, a graph-based pangenomic framework designed to address these challenges. We first demonstrate its application on a comprehensive dataset of Pseudomonas aeruginosa genomes and compare its performance to established genome-scale tools such as PADLOC and DefenseFinder. We then extend this analysis to a broader dataset comprising four enterobacterial species, showcasing PANORAMA's capacity to annotate, compare, and explore the distribution and diversity of phage defense systems across entire genera. Overall, this work provides a powerful new resource for the comparative study of bacterial genomes and highlights the value of pangenomic approaches in revealing the evolutionary dynamics and ecological significance of bacterial defense repertoires.

1 Results and discussion

1.1 Overview of PANORAMA

To predict macromolecular systems, PANORAMA employs rule-based models similar to those used in MacSyFinder [21]. However, instead of applying these models to individual genomes, PANORAMA operates on the pangenome graph structure of PPanGGOLiN. The rules rely on the presence/absence of specific functions predicted from pangenome gene families, incorporating constraints on their genomic organization (i.e. gene colocalization). Functional annotation of gene families of the pangenome graph is performed through alignments with HMM protein profiles [22] defined for each macromolecular system. The genomic contiguity of gene families potentially involved in a system is then assessed on the pangenome graph by applying transitive closure and edge filtering. At the end, the predicted systems consist of sets of colocalized gene families from the pangenome graph, supplemented with information on their classification within the *persistent*, *shell*, or *cloud* genome, as well as their association with RGPs, modules, and spots of integration. Systems are also projected onto the genomes to determine their presence and gene content in each strain. An additional functionality of PANORAMA is its ability to compare pangenomes, identifying similar systems and insertion spots across species. Based on a set of predicted spots or systems in several pangenomes, PANORAMA computes a Gene Family Repertoire Relatedness score for each pair of elements by detecting shared gene families. It then applies a community clustering algorithm to group similar systems or insertion spots into clusters. More details about PANORAMA methods are provided in the Materials and Methods section.

PANORAMA is available as an open-source software written in Python and designed for easy installation to facilitate broader adoption by the community (https://github.com/labgem/PANORAMA). Software commands are organized into two main workflows (Fig. 1). The PanSystem workflow begins by annotating gene families of the pangenome graph using the specified HMM library, then applies system prediction rules from the model repository. Gene family annotations can also be performed externally and provided to PANORAMA by the user in a Tab Separated Values (TSV) file. The PanCompare workflow performs comparative analyses of two or more pangenomes, including gene family clustering and system/spot comparisons. Both workflows generate textual outputs (TSV files), graph-based representations (in GEXF or GraphML formats, compatible with Gephi or Cytoscape software for visualization), and figures to summarize results. Functional annotations and predicted systems are saved in the pangenome's HDF5 file, allowing further analyses. Additional utilities are provided to automatically convert system models and HMM libraries into the PANORAMA format, with support for models from MacSyFinder[9], DefenseFinder[11], CasFinder[23] and PADLOC[10]. Models are stored in JavaScript Object Notation (JSON) format with a flexible and easy-to-understand grammar, enabling users to customize or create new models.

1.2 System prediction benchmark

A set of 941 complete genomes of *Pseudomonas aeruginosa* was used to evaluate PANORAMA's defense system predictions against the reference tools, Defense Finder (including CasFinder models) and PADLOC. Although these two tools use a similar approach to predict defense systems, they differ in the number of models (281 in Defense Finder vs. 385 in PADLOC) and in the parameters used for predicting functions from HMM alignments, as well as for applying presence/absence and colocalization rules. Thanks to its generic system representation, PANORAMA is compatible with both tools and was run using their respective system models and HMMs after format conversion (i.e., PanSystem workflow).

To conduct this benchmark, we assessed whether PANORAMA correctly assigned pangenome gene families to the appropriate systems, based on the results from Defense Finder or PADLOC. As expected, we obtained highly similar results, achieving an F1-score of 99.11% (recall: 99.31%, precision: 98.91%) using PADLOC as a reference and 98.50% (recall: 99.71%, precision: 97.32%) with Defense Finder. As shown in Fig. 2a, a substantial number of families are shared exclusively between PANORAMA and either DefenseFinder (653 families) or PADLOC (879 families), while only 985 families are common.

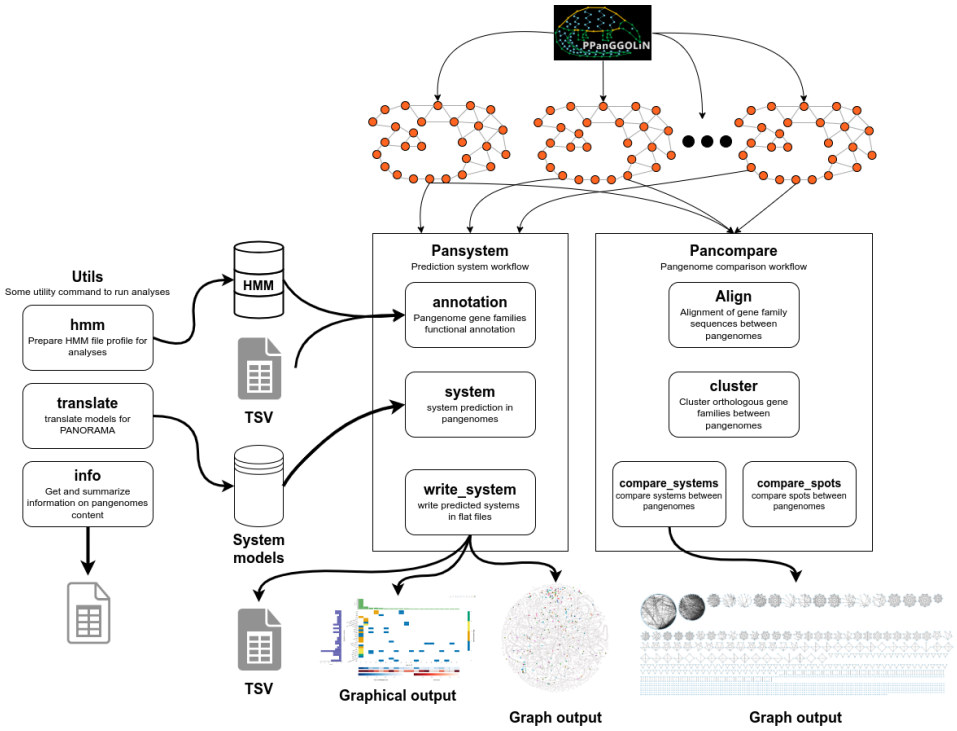

Fig. 1. PANORAMA software overview.Each rounded box represents a possible software command integrated into workflows depicted in square boxes. PANORAMA is organized into two main workflows: *Pansystem*, which focuses on system prediction and annotation within pangenomes, and *Pancompare*, which handles comparative analyses of pangenomes, including gene family alignment, clustering, and system/spot comparisons. Utility tools, such as HMM preparation and information summarization, facilitate data input and model translation. The software outputs include TSV files, graphical visualizations, and graph-based representations for comparing systems and pangenomes.

This highlights the complementary nature of these tools and demonstrates that the pangenome analysis provided by PANORAMA is highly valuable for reconciling their results. Besides, PANORAMA missed only a small number of systems, 25 for Defense Finder and 37 for PADLOC (Table 1). These cases correspond to missing annotations in the pangenome gene families. Unlike PADLOC and Defense Finder, which analyze each sequence individually for every genome, PANORAMA relies on the protein sequence of the family's representative gene for HMM alignments. Consequently, in rare instances, the alignment of the representative sequence falls just below the defined threshold, even though other sequences in the family have hits above it. Conversely, PANORAMA can identify additional results for certain genomes whose gene sequence alignment with the HMM is less conserved than with the representative sequence. A last point concerns the systems predicted only by PANORAMA (i.e., 147 with Defense Finder models and 44 with the ones of PADLOC). Although such a high number was not initially expected, a detailed analysis revealed that some of these additional systems arise because PANORAMA's genomic context search, based on colocalization rules in the pangenome graph, is less stringent than that of PADLOC and DefenseFinder. This relaxed approach allows PANORAMA to detect additional genes that are consistently conserved alongside known system components (see 'Materials and Methods' for further explanation). Other additional predictions stem from differences in HMM annotation due to the use of representative sequences for families, as mentioned earlier, but also from a bug in Defense Finder that affected only RM and Retron systems, where alignment thresholds were not properly handled by the software. This bug has since been corrected by the authors, though it was not tested in this study. Other discrepancies arise because PANORAMA allows multiple associations between gene families and systems, whereas PADLOC and Defense Finder link each gene to only one system. This is especially noticeable for systems with closely related models; for example, PANORAMA often predicts multiple Retron, RM, or Lamassu systems when other tools identify only one (Fig. 2b). One potential improvement would be to include *forbidden* families in the models to facilitate their differentiation, or to implement a scoring function, as done in MacSyFinder, to identify the best system. PANORAMA may also predict additional systems at the genome level for which the colocalization rule is not respected. These systems, flagged as *split*, have their genes separated, possibly due to rearrangements, insertion events, or assembly breaks occurring in a subset of genomes where they are predicted.

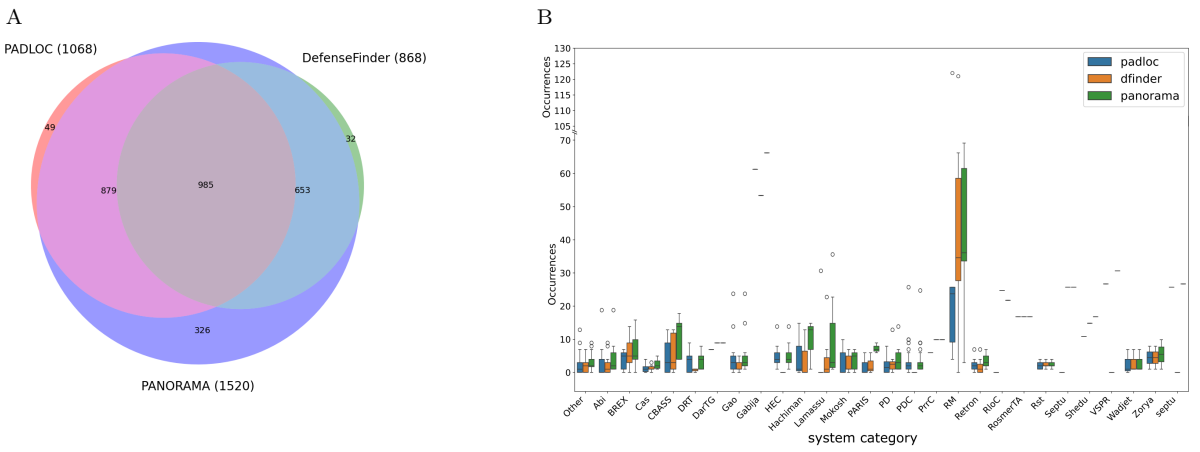

Fig. 2. Comparison of PANORAMA, PADLOC, and DefenseFinder system predictions at the pangenome level. PADLOC and Defense Finder predictions at the genome level were unified at the pangenome level by converting sets of system genes to sets of gene families. (A) Venn diagram illustrating the overlap of system families predicted by the three tools. (B) Boxplots displaying the distribution of system counts for each category, as predicted by the different tools.

Table 1. System prediction comparison between PANORAMA, PADLOC and DefenseFinder

Method 1 / Method 2	Common M1	Common M2	% common	Specific M1	Specific M2
PADLOC (1068) / DFinder (868)	502	506	52,07 %	566	362
PADLOC (1068) / PANORAMA (1064)	1031	1020	96,20 %	37	44
DFinder (868) / PANORAMA (976)	843	829	90,67 %	25	147

The tools were also evaluated for execution performance by measuring their runtime, CPU time and memory usage on a Linux server with 36 CPU cores (Table 2). Since PADLOC and Defense Finder are not designed to handle multiple genomes or parallelize computations, individual commands were executed for each genome, distributing the workload across all available CPU cores. PANORAMA significantly outperforms the other tools in runtime, being 3 to 10 times faster, while exhibiting similar memory usage. This efficiency is achieved by analyzing gene families rather than individual genes, which reduces computational overhead. Additionally, PANORAMA utilizes pyHMMER [24] for HMM alignments, optimizing the workflow by minimizing I/O operations and enabling on-the-fly result filtering. In contrast, other tools use the HMMER software directly [22], which necessitates post-processing steps.

Table 2. Benchmark results.

Tool (version)	Database (version)	#Systems predicted	Run Time (h)	CPU Time (h)	Peak Memory (GB)
DefenseFinder (1.2.2)	Defense-finder-models (1.2.4) & CasFinder (3.1.0)	881	1.55	29.02	6.84
PANORAMA (1.0.0)		976	0.51	0.82	11.22
PADLOC (2.0.0)	PADLOC-DB (2.0.0)	1090	2.58	88.72	9.24
PANORAMA (1.0.0)		1064	0.24	0.78	13.05

1.3 *Pseudomonas aeruginosa* defense system analysis

1.3.1 System prediction and analysis

Using the same set of *P. aeruginosa* genomes as for the benchmark, PANORAMA's defense system predictions with Defense Finder models were analyzed in greater detail. PANORAMA identified a total of 976 systems in the pangenome from 154 distinct models, with restriction-modification (RM) systems being the most abundant (Fig. 3). RM are present in 84% of genomes, with nearly 300 systems detected, of which type I systems are the most common, occurring 130 times. The Gabija, CRISPR-Cas, and CBASS system categories follow as the next most prevalent defense systems in genomes, each with a presence rate above 40%. These observations corroborate the study of Johnson *et al.* [25]. At the pangenome level, some system categories are highly prevalent across genomes but are represented by only a few distinct systems. For example, CRISPR-Cas systems appear in 48% of genomes, yet only 13 distinct systems are identified in the pangenome. This is further highlighted by a Shannon entropy calculation, which measures the compositional diversity of system categories (Fig. 3c). For CRISPR-Cas systems, the entropy is 1.35, indicating a high degree of conservation in their gene family composition across genomes. Among the most prevalent system categories, others show notable diversity, including RM, Gabija, and RloC. The RloC systems, for example, consist of 22 distinct systems spread across 35% of genomes, with a Shannon entropy of 21.58, indicating considerable variability in their gene family composition. These predictions are consistent with recent studies and highlight the remarkable diversity of anti-phage immune systems in prokaryotes [11].

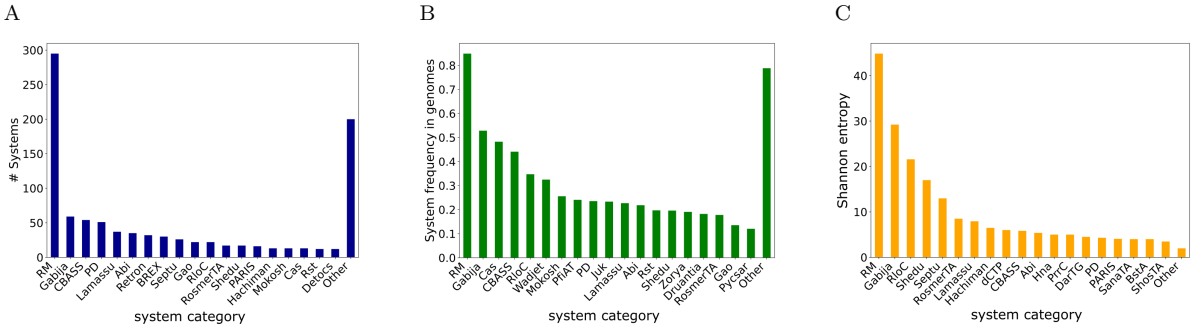

Fig. 3. System prediction metrics in *P. aeruginosa*. Systems are grouped by categories on the x-axis and ordered by decreasing values. Only the 19 highest-value system categories are displayed, with others grouped under the "Other" category. (A) Number of systems found for each category in the pangenome. (B) Relative frequency of system categories in genomes. (C) Shannon entropy of system categories.

1.3.2 Defense islands and spots of insertion

PANORAMA systems can be analyzed in conjunction with additional information extracted from the PPanGGOLiN pangenome graph, particularly concerning their association with the variable genome and their localization within RGPs and insertion spots predicted by panRGP [7]. This enables the identification of defense islands (i.e., variable regions enriched with defense systems) and their hotspots (i.e., frequently occurring insertion sites of defense islands in genomes).

Most defense systems are predicted within the variable (*shell* or *cloud*) genome of *P. aeruginosa* and are located in spots. PANORAMA identified 247 spots containing at least one defense system, representing 25% of all pangenome spots. Among them, 4 spots (7, 6, 45, 69) have a high frequency (¿25%) and exhibit the highest number of defense systems predicted at the pangenome level (¿60 systems) (Fig. 4). Notably, spots 7 and 6 are the most diverse, harboring 238 and 162 associated systems, respectively Fig. 5. They are mostly composed of RM systems (51% and 56%) but also exhibit a broad diversity of other categories, including BREX (6%), Gabija (5% and 4%), PD (4% in spot 7) and CBASS (4% in spot 6). These two spots were previously identified using a non-automated approach in the study by Johnson *et al.* [25] as core defense hotspots in *P. aeruginosa*, where they were designated CDHS-1 and CDHS-2. This further highlights the value and reliability of PANORAMA in automatically detecting defense islands and their insertion spots. Using PANORAMA, we also identified two additional defense hotspots (spots 45 and 69). Spot 45 contains 110 systems and stands out as the most balanced in terms of system categories; it is also the only hotspot with a notable presence of PARIS systems (8%). Spot 69, like spots 7 and 6, is dominated by RM systems, with PrrC systems specifically represented at 7%. Although less frequently observed across genomes (¡20%), spots 61 and 1 display highly diversified system content, comprising 72 and 65 systems, respectively. Both are also rich in RM systems, with Mokosh particularly represented in spot 1 and CBASS and PD systems notably present in spot 61 (8%). Finally, spots 4 and 9 are relatively frequent across genomes (30%) but contain few distinct systems, 31 and 10, respectively. These results highlight the potential of PANORAMA to provide a comprehensive landscape of defense systems in a species, enabling pangenome-scale analysis and the identification of defense islands with their hotspots of integration.

Fig. 4. System diversity and defense spot frequency in *P. aeruginosa*. This bubble plot displays the distribution of defense spots identified by PANORAMA, based on their frequency in genomes (x-axis) and the total number of defense systems identified within each spot at the pangenome level (y-axis). The size of the bubbles is proportional to the number of distinct system categories represented in each spot (legend at top shows scale).

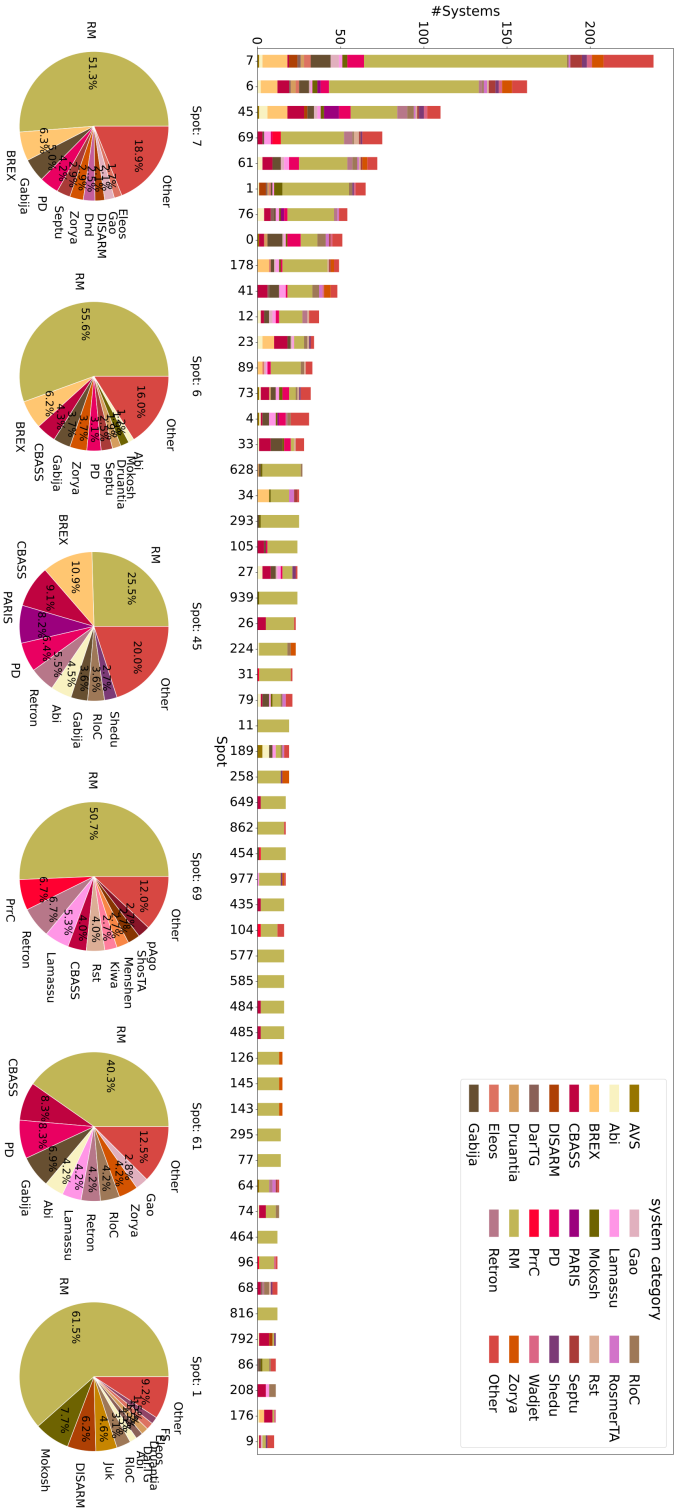

Fig. 5. Defense systems within insertion spots of the *P. aeruginosa* pangenome. The bar plot (top) displays the number of predicted defense systems in the *P. aeruginosa* pangenome for each insertion spot. Only spots with at least 10 systems are displayed. The pie charts (bottom) illustrate the system category composition for the six major insertion spots.

1.4 Pangenome comparison of Enterobacteriaceae defense arsenal

To demonstrate the comparative functionalities of PANORAMA (i.e., PanCompare workflow), the pangenomes of four Enterobacteriaceae species were analyzed for defense system and spot prediction. This dataset represent more than 6,000 genomes from *Citrobacter freundii*, *Salmonella enterica*, *Klebsiella pneumoniae* and *Escherichia coli* species (Table 3). The distribution of systems between species and their association with conserved spots were studied. Defense systems were predicted using Defense Finder models.

1.4.1 Defense system distribution in the four species

A total of 351, 461, 1005 and 1448 defense systems were predicted from the pangenomes of *Citrobacter freundii*, *Salmonella enterica*, *Klebsiella pneumoniae* and *Escherichia coli*, respectively. In addition to textual outputs, PANORAMA automatically generates a heatmap that displays system occurrences across the compared pangenomes (see Fig. S1). Among the different categories, RM systems are the most abundant in the four species, accounting for 30% to 45% of the systems found. Following RM systems, PD systems are the next most prevalent, representing 5% to 7% of the systems within Enterobacteriaceae (Fig. 6). Other notable system categories, such as Retron, CBASS, and Abi, are also relatively abundant across all pangenomes. Next, we evaluated the species-specificity of each system category by computing enrichment factors (Fig. S2). Our findings indicate that certain categories, Abi and Dnd in *S. enterica*, Juk and pAgo in *K. pneumoniae*, and Bunzi and RADAR in *C. freundii*, exhibit enrichment factors above 3, highlighting their preferential association with these species. Tiamat systems are only found in *S. enterica* and *C. freundii*. Interestingly, *E. coli* does not show any systems in higher abundance compared to other species, a sign of a more balanced diversity in its defense mechanisms.

1.4.2 Identification of conserved spots

With PANORAMA, we searched for similar spots based on their related gene families, using a gene family repertoire relatedness (GFRR, see Section subsection 3.2) threshold of at least 60%. This threshold guarantees at least one similar gene family on each side of the border. As shown in Fig. 7, we identified 99 clusters of similar spots, corresponding to 219 spots that have at least one spot with a similar bordering gene family composition with another one in an *Enterobacteriaceae* pangenomes.

E. coli is the species that shares the most spots with others (151), followed by *S. enterica* (131), then *C. freundii* (112) and *K. pneumoniae* (104). Proportionally to its number of spots, *C. freundii* is the species with the most similar spots, with 35% of its spots similar to the other pangenomes. The 2 species with the most common spots are *E. coli* and *S. enterica*, with 80 common spots. PANORAMA can then be used to highlight insertion spots at a higher taxonomic rank than the species. These could be areas of interest for research into shared evolution or exchanges between these species.

1.4.3 Spot identification and conservation

Using the comparative functionalities of PANORAMA, we identified 99 clusters of similar spots conserved in at least two of the four Enterobacteriaceae species (Fig. 7). As might be expected given their phylogenetic proximity, *E. coli* and *S. enterica* share the most common spots (i.e., 34 spot clusters, 25 of which are found only in these two species). About half of the spot clusters (n=47) are associated with a defense system in at least one species, comprising a total of 520 distinct defense systems among the 3,265 identified in the four species pangenomes. Of these, there is only one spot cluster (cluster 58) conserved in all pangenomes containing 34 defense systems. *E. coli* and *S. enterica* harbor the highest number of defense systems (263 systems) across their specific spot clusters (7 clusters). One of these clusters (spot cluster 409) includes spot 86 in *S. enterica* and spot 175 in *E. coli*. These spots rank sixth in terms of the number of defense systems, with 35 in *S. enterica* and 218 in *E. coli*, and can therefore be considered as potential defense hotspots. Analyzing their composition reveals notable similarities (Fig. S3). Both spots are mainly composed of RM systems (55% in *S. enterica*, 85% in *E. coli*), followed by BREX, CBASS,

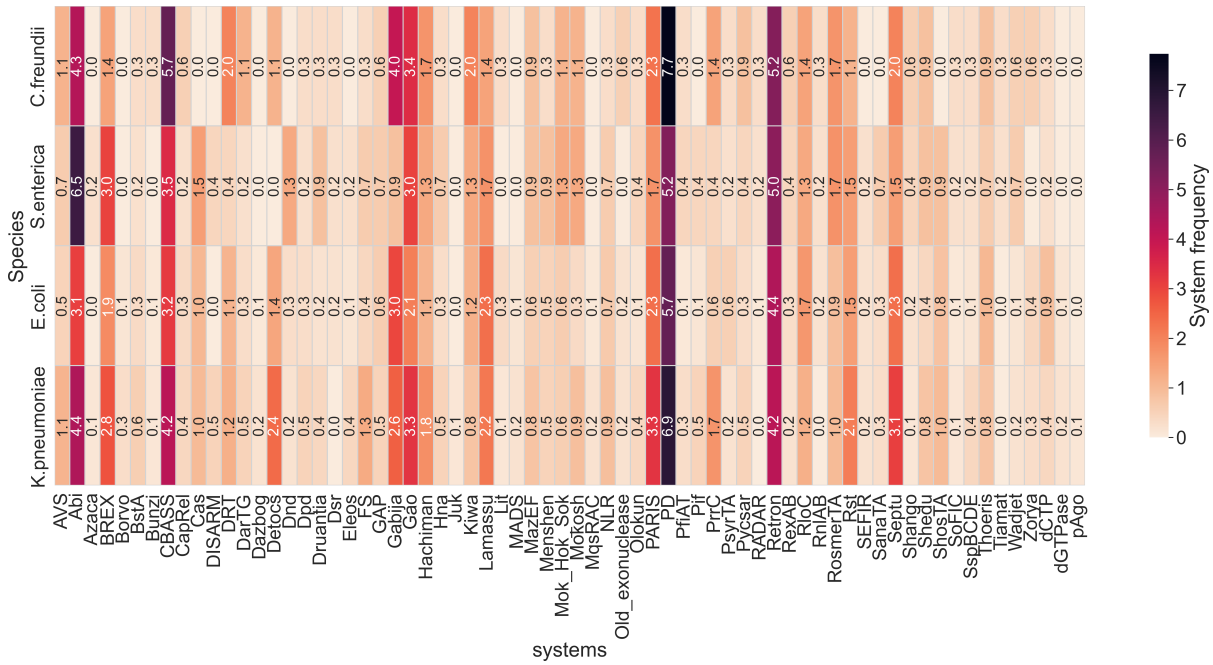

Fig. 6. Relative frequency of system categories in Enterobacteriaceae pangenomes. The relative frequencies are expressed as a percentage. RM system category was removed to get a clearer view.

and PrrC, which are generally not phage-specific. These findings support the hypothesis that defense systems may have been exchanged between the two species from this conserved hotspot. Examining the system category diversity of other spot clusters (Fig. 8), many clusters are predominantly composed of RM systems. Some clusters are dedicated to a single system category, such as BstA in cluster 2320, while others, like the previously mentioned cluster 58, are more diverse, bringing together systems from all four studied species.

Beyond their illustrative purpose, the analyses presented here highlight the ability of PANORAMA's comparative functionalities to identify conserved defense islands across species, providing valuable insights into the evolution of defense systems and their mechanism of acquisition.

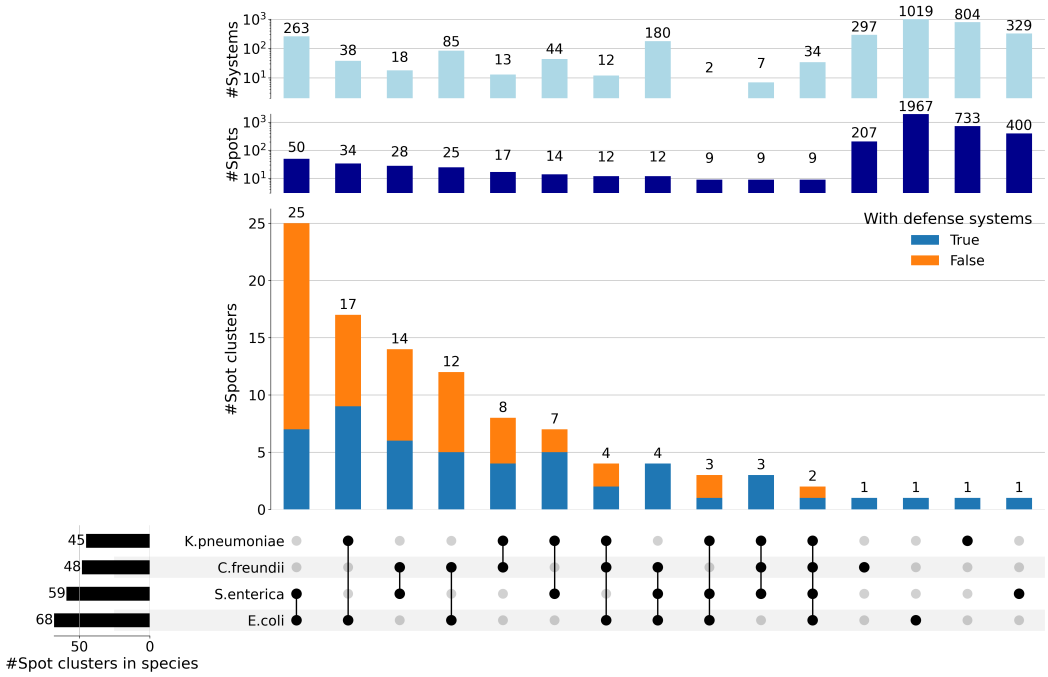

Fig. 7. Sharing of spot clusters across four species and their association with defense systems. The UpSet plot shows the number of spot clusters shared across the four compared species with stacked bars to indicate whether they contain defense systems (in blue) or not (in orange). The two top bar plots represent spot and defense system abundance metrics on a logarithmic scale.

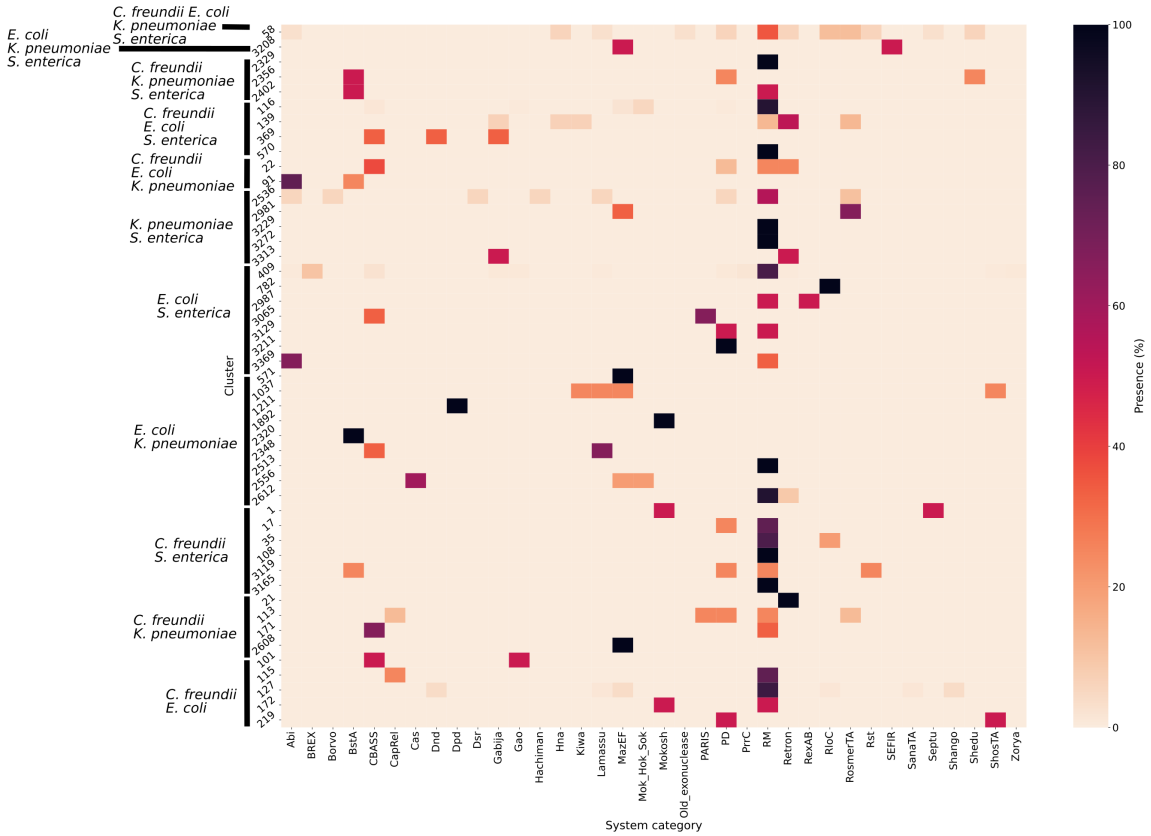

Fig. 8. Relative frequency of system categories in each spot cluster. The heatmap presents the relative frequency (%) of system categories per spot cluster. Species associated with each spot cluster are listed next to the cluster identifiers.

2 Conclusion

In this work, we introduced PANORAMA, a novel open-source framework for the prediction and analysis of macromolecular systems across prokaryotic pangenomes. Unlike existing tools that operate on individual genomes, PANORAMA leverages the pangenome graph structure provided by PPanGGOLiN to perform system-level analyses at both species and interspecies scales. It utilizes rule-based models inspired by those in MacSyFinder, but adapted to pangenomes, enabling the flexible and accurate identification of diverse functional systems, such as defense mechanisms.

Benchmarking against established tools like DefenseFinder and PADLOC demonstrated that PANORAMA achieves comparable prediction accuracy while reducing computational demands, making it particularly well-suited for large-scale analyses involving hundreds or thousands of genomes.

Beyond its performance, PANORAMA introduces key methodological innovations. Notably, its graph-based approach enables the detection of conserved genomic contexts by identifying gene families that consistently co-occur across multiple genomes. This strategy allows for the robust identification of evolutionarily maintained system components, even when their genomic proximity is disrupted in some genomes by rearrangements, insertions, or assembly fragmentation. As a result, PANORAMA is more flexible than conventional genome-based tools to detect atypical genomic architectures.

Applied to hundreds of genomes of a species, PANORAMA can quickly identify the different systems and their occurrence in individual genomes. A notable strength lies in its ability to associate systems with regions of genomic plasticity and insertion hotspots, allowing the identification of genomic islands enriched with systems and their preferred integration sites. In *P. aeruginosa*, four major hotspots were identified, including two that correspond to previously published core defense hotspots.

To further illustrate the comparative functionality of PANORAMA, we analyzed the defense repertoires of over 6,000 genomes from four Enterobacteriaceae species. PANORAMA revealed both conserved and species-specific system categories and was able to cluster insertion spots based on shared gene family content. This enabled the identification of conserved defense islands across phylogenetically related species, offering insights into the evolutionary conservation of these systems.

Altogether, PANORAMA provides a robust and extensible framework for system detection and comparative analysis at the pangenome level. Its flexible modeling format, compatibility with existing system model databases, and integrative workflows make it a valuable tool for microbiologists, bioinformaticians, and evolutionary biologists interested in understanding the distribution, function, and evolution of macromolecular systems. As the number of available genomes continues to grow, tools like PANORAMA will become increasingly critical in unveiling the complex architectures of microbial defense and other macromolecular systems.

Looking ahead, we plan to incorporate additional rule-based approaches to predict metabolic modules, which consist of the detection of genomic contexts encoding enzymes involved in the same pathway, leveraging pathway definitions from comprehensive databases such as KEGG [26] or MetaCyc [27]. We also aim to develop specific rules for detecting clusters of Carbohydrate-Active Enzymes (CAZymes) involved in polysaccharide biosynthesis and degradation. Furthermore, the pangenome graph framework provided by PANORAMA offers a promising avenue to explore genomic context and uncover novel systems. By incorporating fuzzy functional predictors, based on structural similarity or protein language models, we hope to detect new systems made of functionally analogous components co-localized within the pangenome graph.

3 Materials and Methods

3.1 Pangenome system detection workflow

To predict macromolecular systems, PANORAMA employs rule-based models that analyze the presence/absence of specific functions predicted from pangenome gene families while considering constraints on their genomic organization. The PanSystem workflow begins with gene family annotation using HMM libraries, followed by the identification and evaluation of genomic contexts within the pangenome graph based on system model rules (Fig. 9). Finally, the predicted systems at the pangenome level are mapped onto individual genomes to assess their presence and gene content.

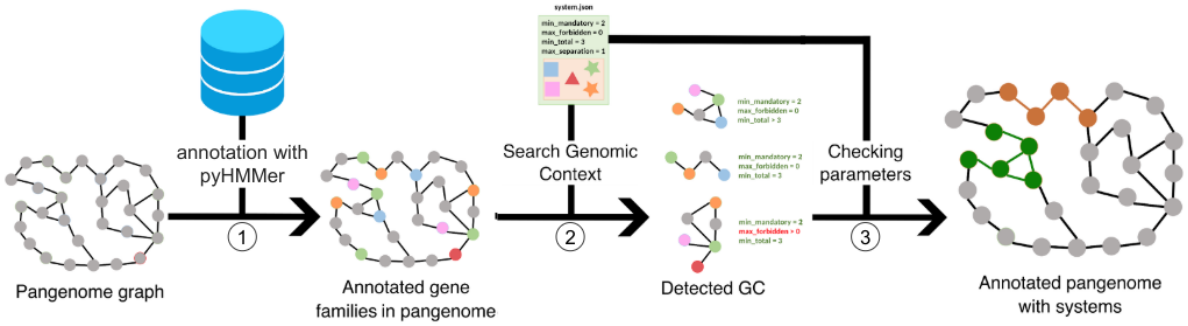

Fig. 9. PANORAMA PanSystem workflow. The detection workflow for macromolecular systems consists of multiple sequential steps: (1) gene family annotation using HMM profiles, (2) detection of genomic contexts from the pangenome graph containing annotated families, (3) checking model rules, (4) projection of systems on genomes.

3.1.1 System modeling

The Models used by PANORAMA for predicting macromolecular systems are similar to those of MacSyFinder [9] but differ in some aspects (Fig. 10). The primary components of system Models are Families (i.e. isofunctional protein families) instead of genes, and an additional hierarchical level is introduced to represent Functional Units. A Functional Unit is defined as a set of Families that work together to perform a necessary function for the system. Several Functional Units may be required for a system to operate effectively. This new level provides a more detailed and accurate description of systems, allowing the use of distinct rules for presence/absence and genomic organization both for the Families of a Functional Unit and between Functional Units. In PANORAMA, we also introduce the concept of canonical Models, which represent a more relaxed definition of systems. In PADLOC [10], these models are identified by the keyword "other" in their name. While these systems may not have been experimentally validated and might not be functional, their prediction can be valuable for identifying new systems or potential variants. During system detection, priority is given to non-canonical Models. A canonical Model is predicted only if its Families are not already associated with a non-canonical Model.

For presence/absence constraints, each Functional Unit and Family are categorized as *mandatory*, *accessory*, *neutral*, or *forbidden*. *Mandatory* elements are essential for the system. *Accessory* components are dispensable. They contribute to the system but may not be identified in all system variants due to rapid evolution or the absence of homology. *Forbidden* elements are incompatible with system functionality. They can help differentiate systems with shared components or distinguish inhibited systems. *Neutral* components are considered as associated functions but are not used to assess system predictions. However, they can serve as intermediaries in genomic context analysis by linking mandatory or accessory elements. To predict a system, quorum rules on component presence/absence are defined by two parameters: *minimum_mandatory* (the minimum number of mandatory elements required) and *minimum_total* (the minimum number of both mandatory and accessory elements). These parameters are specified at two levels: at the Model level, to assess the presence of Functional Units, and at the Functional Unit level, to evaluate predicted Families. Constraints on the genomic organization of a

system are governed by a *transitivity* parameter, which defines the maximum genomic distance between gene families in the pangenome graph corresponding to the system components (see 'Pangenomic context extraction' section).

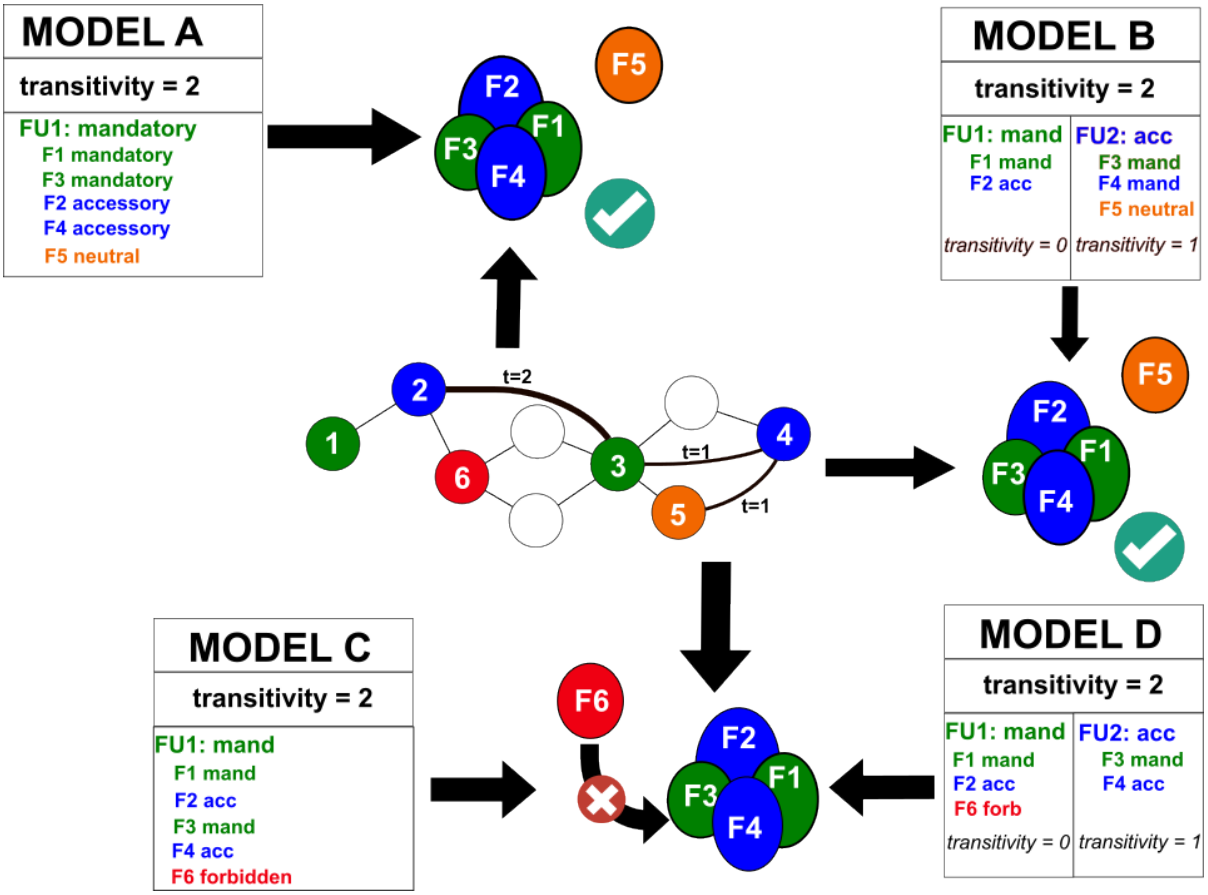

Fig. 10. PANORAMA system Modeling. Four system Models, named A, B, C, and D, are presented as toy examples to illustrate model rules. At the center, the corresponding genomic context extracted from the pangenome graph is shown. Nodes representing target families (i.e., model families) are labeled with Family numbers, and edges between them are bold, with dotted lines if they are from a transitive closure ($t >= 1$). Only transitivity edges relevant for model evaluation are represented. Functional Units (FU) and Families (F) are color-coded by their category: green for *mandatory*, blue for *accessory*, orange for *neutral*, and red for *forbidden*. Only models A (one FU) and B (two FU) are predicted. Others (C and D) are not predicted, as their extracted genomic context contains a *forbidden* Family.

3.1.2 Functional annotation of pangenome gene families

To annotate the pangenome graph, PANORAMA utilizes HMM libraries in which each HMM represents a specific Family defined in the system Models. Protein sequences of the family's representative genes are aligned to the HMM profiles using the hmmsearch method from the pyHMMER Python library [24]. PANORAMA also offers the ability to use the consensus protein sequence of gene families. It is determined by performing a multiple sequence alignment (MSA) excluding fragmented genes and then computing the consensus sequence with pyHMMER. To validate the alignments, a metadata file can be linked to an HMM library, specifying thresholds for each HMM based on various criteria, such as alignment coverage on the sequence or profile, score, e-value, and independent e-value. Alternatively, a global threshold can be applied to all HMMs. An option is also available to keep only the n best hits. Gene family annotation can also be performed externally using alternative prediction methods. In this case, a TSV file can be provided to describe the functions associated with each gene family.

3.1.3 Pangenome context extraction

For each system, connected components between gene families corresponding to Model Families (hereafter called target families) are searched for in the pangenome graph. According to the model rules on genomic colocalization of system components (i.e. the *transitivity* parameter), additional edges are added to the pangenome graph between families separated by less than t genes in the corresponding genomes. This corresponds to a partial transitive closure on the pangenome graph, enabling the connection of two target families even if their genes are not directly adjacent in the genomes. Next, a Jaccard index-based criterion (Equation 1) is computed on edges to evaluate the genomic context conservation (i.e. synteny conservation). For two families a and b connected by an edge $e_{a,b}$, the index $J_{a,b}$ is defined as the ratio of the number of genomes where the edge $e_{a,b}$ exists to the number of genomes in which at least one of the two gene families is present ($|\text{gen}_a \cup \text{gen}_b|$). To enhance the detection of systems present in a limited number of genomes, this index is computed locally, considering only genomes that contain the target families of the connected component rather than all the genomes of the pangenome. Edges having a Jaccard index below a defined threshold ($J_{a,b} < 0.8$ by default) are removed. The connected components with their remaining nodes are then evaluated in terms of presence/absence rules to validate system prediction. For each connected component, this process, which combines edge filtering and presence/absence rule validation, is applied iteratively, starting with the largest set of target families observed in a genome and then exploring other sets of target families, from largest to smallest, if they are not already included in a predicted system.

$$J_{a,b} = \frac{|\text{gen}_{e_{a,b}}|}{|\text{gen}_a \cup \text{gen}_b|} \tag{1}$$

Finally, the connected components corresponding to predicted systems at the pangenome level are mapped back onto individual genomes. Their occurrence in individual strains is assessed based on presence/absence and colocalization rules. They are classified in 3 distinct categories: *strict*, *extended* or *split*. A system is flagged as *strict* if the *transitivity* parameter is respected between genes belonging to the system. If additional genes of the connected component are found between those encoding the system, it is classified as *extended*. Indeed, applying transitive closure to the pangenome graph enables the detection of additional genes conserved with those of the system. This provides a more relaxed definition of colocalization rules than the strict intergenic space constraints employed by MacSyFinder or PADLOC. Lastly, *split* systems are those where system genes are separated by non-member genes of the connected component detected at the pangenome level or are located on different contigs. Such fragmentation can occur in a subset of genomes due to rearrangements, insertion events (e.g., insertion sequences), or assembly breaks.

3.2 Pangenome comparison workflow

The PanCompare workflow of PANORAMA enables the comparison of pangenomes and the identification of similar elements across species, such as macromolecular systems or spots of insertion. These elements are represented as sets of gene families. The first step consists in clustering all the gene families from the different pangenomes to identify groups of homologous gene families. Then, a Gene Family Repertoire Relatedness score for each pair of elements is computed. Finally, a community clustering algorithm is applied to identify clusters of elements sharing similar gene family content.

3.2.1 Gene families clustering

From all pangenomes to compare, representative protein sequences of gene families are extracted and clustered using MMSeqs2 cluster command [28] to obtain groups of homologous gene families (GGFs) sharing at least 50% of amino acid identity (–min-seq-id parameter) with an alignment coverage of 80% (-c parameter) by default. The following additional parameters are used: *–max-seqs* 400 *–min-ungapped-score* 1 *–kmer-per-seq* 80 *–alignment-mode* 2 *–cluster-mode* 1.

3.2.2 Gene Family Repertoire Relatedness score

To compare the family content of two pangenome elements, two Gene Family Repertoire Relatedness (GFRR) scores are computed using GGFs. For two pangenome elements, a and b, with their GFG content denoted as GFG_a and GFG_b, the minimal GFRR score, $minGFRR_{a,b}$, is defined as the ratio of the number of common GGFs to the minimum number of GGFs between a and b (Equation 2). Similarly, the maximal GFRR score, $maxGFRR_{a,b}$, is computed using the maximum number of GGFs (Equation 3).

$$minGFRR_{a,b} = \frac{|\mathrm{GFG}_a \cap \mathrm{GFG}_b|}{\min(|\mathrm{GFG}_a|, |\mathrm{GFG}_b|)} \tag{2}$$

$$maxGFRR_{a,b} = \frac{|\mathrm{GFG}_a \cap \mathrm{GFG}_b|}{\max(|\mathrm{GFG}_a|, |\mathrm{GFG}_b|)} \tag{3}$$

GFRR score computation can be applied to predicted macromolecular systems or spots of integrations. For spots, the two sets of gene families corresponding to spot borders are merged to compute the GFRR scores.

3.2.3 Pangenome element clustering

To identify similar elements (i.e. spots of insertion or systems) between pangenomes, GFRR scores are computed for each pair of elements. A graph is then constructed, where nodes represent pangenome elements, and edges are weighted by their corresponding GFRR scores. Edges are filtered according to a GFRR threshold ($minGFRR >= 0.6$ by default) to ensure strong similarity between connected elements. Finally, a Louvain algorithm [29], using NetworkX library implementation with GFRR weight on edges, is applied to the graph to identify non-overlapping communities corresponding to groups of similar elements between pangenomes. This functionality was used to identify conserved spots of insertion between species containing defense systems. A system is associated with a spot if all of its gene families are part of RGPs found within the boundaries of that spot.

3.3 Data, benchmark, and metrics

3.3.1 Genomic data and defense system prediction

Complete genomes of five species were downloaded from NCBI RefSeq [30] and analyzed for defense system prediction using Defense Finder (v1.2.2 with 239 models of v1.2.4 + 43 CasFinder models of v3.1.0), PADLOC (v2.0.0, 385 models of v2.0.0) and PANORAMA (v1.0.0). The dataset includes *Pseudomonas aeruginosa* (941 genomes) and four well-studied Enterobacteriaceae species: *Citrobacter freundii* (79 genomes), *Escherichia coli* (3,083 genomes), *Klebsiella pneumoniae* (1,659 genomes) and *Salmonella enterica* (1,380 genomes). Before running PANORAMA, the pangenomes of the five species were obtained using PPanGGOLiN v2.1.2 with default parameters and keeping the original RefSeq annotations. Pangenome metrics, including the number of families categorized as persistent, shell or cloud genomes, as well as the number of predicted RGPs and spots of insertion, are summarized in Table 3. These metrics are also available through the info command of PANORAMA. Models from Defense Finder and PADLOC (n=667), along with their associated HMMs (n=6,272), were converted in PANORAMA format using its internal conversion utility.

Table 3. Pangenomes content

Species	Genomes	Genes	Families	Edges	Persistent	Shell	Cloud	RGPs	Spots
C. freundii	79	385611	16865	24936	3780	2954	10131	3863	254
E. coli	3083	14447407	44278	140743	3018	7174	34086	240163	2036
K. pneumoniae	1659	8851686	32685	75073	4155	5085	23445	67678	778
P. aeruginosa	941	5811655	34058	61841	4925	6768	22365	44935	984
S. enterica	1380	6222501	23941	46535	3474	4080	16387	65920	458

3.3.2 Benchmark protocol

The *P. aeruginosa* genome dataset was used to evaluate PANORAMA's defense system predictions against the reference tools, Defense Finder v1.2.2 and PADLOC v2.0.0, using their respective system models and HMMs. To ensure consistency in comparison, the predictions from Defense Finder and PADLOC, originally consisting of gene sets associated with defense systems, were mapped to pangenome gene families by linking each identified gene to its corresponding family. Then, we define a True Positive (TP) as a gene family assigned to the same defense system by both PANORAMA and the reference tool. Gene families associated with a system by the reference tool but not predicted by PANORAMA are classified as False Negatives (FN). Gene families assigned to a different system or not predicted by the reference tool are considered False Positives (FP). To assess the performance of PANORAMA, we computed precision, recall, and F1-score using the following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4a}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4b}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4c}$$

The benchmark was conducted on a Linux server equipped with two Intel® Xeon® Gold 6150 processors (36 cores), 376 GiB of DDR4 RAM, running CentOS v7.9.2009 with kernel 3.10.0 and Python 3.10. Since Defense Finder and PADLOC cannot process multiple genomes simultaneously and parallelize computations, individual commands were executed for each genome, distributing the workload across the 36 available cores without concurrency. PANORAMA was run with the –threads parameter set to 36 to enable parallel computation. To evaluate execution performance, total run time, CPU time, and peak memory usage were measured for each tool. For consistency in comparison, the pangenome construction step made by PPanGGOLiN was not included in PANORAMA execution metrics as pangenomes are inputs of PANORAMA, like genomes for PADLOC and Defense Finder.

3.3.3 System category diversity

To assess the compositional diversity of systems within a given category (e.g., Restriction Modification, Gabija, CRISPR-Cas) in a pangenome, a Shannon entropy is calculated as follows:

$$H(C, p) = -\sum_{i \in C} P(sm_i) \log_2 P(sm_i) \tag{5}$$

Where:

- $H(C, p)$: the Shannon entropy of a system category C in a pangenome p.
- $P(sm_i)$: probability of a system model sm_i of C to occur in p, calculated for each system model as the ratio of the number of gene families of p associated with sm_i to the total number of gene families across all systems of C.

The higher the entropy, the more diversified the gene families that make up the systems in a given category.

3.3.4 System category enrichment

To determine the specificity of a system category in a set of pangenomes, an enrichment factor is computed as follows:

$$EF(C,p) = \frac{f_r(C,p)}{f_r(C,P)} \quad (6)$$

Where:

- C: a system category.
- P: a collection of pangenomes p.
- $f_r(C,p)$: the relative frequency of the system category C in the pangenome p, calculated as the ratio of the number of systems of C predicted in p to the total number of systems in p.
- $f_r(C,P)$: the relative frequency of the system category C in P, calculated as the ratio of the number of systems of C predicted in all $p \in P$ to the total number of systems across all $p \in P$.

An enrichment factor above 1 indicates that a system category appears n times more in a specific pangenome than in the entire collection of pangenomes.

Data availability

Acknowledgments

References

1. Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. en. *Functional & Integrative Genomics* **15.** Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 2 Publisher: Springer Berlin Heidelberg, 141–161. ISSN: 1438-7948. https://link.springer.com/article/10.1007/s10142-015-0433-4 (2025) (Mar. 2015).
2. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. en. *Nature Biotechnology* **36.** Publisher: Nature Publishing Group, 996–1004. ISSN: 1546-1696. https://www.nature.com/articles/nbt.4229 (2025) (Nov. 2018).
3. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". eng. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 13950–13955. ISSN: 0027-8424 (Sept. 2005).
4. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Current Opinion in Microbiology. Host–microbe interactions: bacteria • Genomics* **23,** 148–154. ISSN: 1369-5274. https://www.sciencedirect.com/science/article/pii/S1369527414001830 (2025) (Feb. 2015).
5. The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics* **19,** 118–135. ISSN: 1477-4054. https://doi.org/10.1093/bib/bbw089 (2024) (Jan. 2018).
6. Gautreau, G. *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. en. *PLOS Computational Biology* **16.** Publisher: Public Library of Science, e1007732. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007732 (2022) (Mar. 2020).
7. Bazin, A., Gautreau, G., Médigue, C., Vallenet, D. & Calteau, A. panRGP: a pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics* **36,** i651–i658. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btaa792 (2022) (Dec. 2020).
8. Bazin, A., Medigue, C., Vallenet, D. & Calteau, A. *panModule: detecting conserved modules in the variable regions of a pangenome graph* en. Tech. rep. Section: New Results Type: article (bioRxiv, Dec. 2021), 2021.12.06.471380. https://www.biorxiv.org/content/10.1101/2021.12.06.471380v1 (2022).

9. Néron, B. *et al.* MacSyFinder v2: Improved modelling and search engine to identify molecular systems in genomes. fr. *Peer Community Journal* **3.** ISSN: 2804-3871. https://peercommunityjournal.org/articles/10.24072/pcjournal.250/ (2024) (2023).
10. Payne, L. J. *et al.* Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Research* **49,** 10868–10878. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkab883 (2022) (Nov. 2021).
11. Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. en. *Nature Communications* **13.** Publisher: Nature Publishing Group, 2561. ISSN: 2041-1723. https://www.nature.com/articles/s41467-022-30269-9 (2024) (May 2022).
12. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* **21,** 180. ISSN: 1474-760X. https://doi.org/10.1186/s13059-020-02090-4 (2024) (July 2020).
13. Jonkheer, E. M. *et al.* PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics* **38,** 4403–4405. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btac506 (2024) (Sept. 2022).
14. Noll, N., Molari, M., Shaw, L. P. & Neher, R. A. PanGraph: scalable bacterial pan-genome graph construction. *Microbial Genomics* **9.** Publisher: Microbiology Society, 001034. ISSN: 2057-5858. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001034 (2024) (2023).
15. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. en. *Nature Reviews Microbiology* **18.** Publisher: Nature Publishing Group, 67–83. ISSN: 1740-1534. https://www.nature.com/articles/s41579-019-0299-x (2025) (Feb. 2020).
16. Goldfarb, T. *et al.* BREX is a novel phage resistance system widespread in microbial genomes. *The EMBO Journal* **34.** Num Pages: 183 Publisher: John Wiley & Sons, Ltd, 169–183. ISSN: 0261-4189. https://www.embopress.org/doi/full/10.15252/embj.201489455 (2025) (Jan. 2015).
17. Ofir, G. *et al.* DISARM is a widespread bacterial defence system with broad anti-phage activities. en. *Nature Microbiology* **3.** Publisher: Nature Publishing Group, 90–98. ISSN: 2058-5276. https://www.nature.com/articles/s41564-017-0051-0 (2025) (Jan. 2018).
18. Millman, A. *et al.* Bacterial Retrons Function In Anti-Phage Defense. eng. *Cell* **183,** 1551–1561.e12. ISSN: 1097-4172 (Dec. 2020).
19. Georjon, H. & Bernheim, A. The highly diverse antiphage defence systems of bacteria. en. *Nature Reviews Microbiology* **21.** Publisher: Nature Publishing Group, 686–700. ISSN: 1740-1534. https://www.nature.com/articles/s41579-023-00934-x (2025) (Oct. 2023).
20. Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolutionary Genomics of Defense Systems in Archaea and Bacteria*. en. *Annual Review of Microbiology* **71.** Publisher: Annual Reviews, 233–261. ISSN: 0066-4227, 1545-3251. https://www.annualreviews.org/content/journals/10.1146/annurev-micro-090816-093830 (2025) (Sept. 2017).
21. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. en. *PLOS ONE* **9.** Publisher: Public Library of Science, e110726. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0110726 (2022) (Oct. 2014).
22. Eddy, S. R. Accelerated Profile HMM Searches. en. *PLOS Computational Biology* **7.** Publisher: Public Library of Science, e1002195. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002195 (2024) (Oct. 2011).
23. Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research* **46,** W246–W251. ISSN: 0305-1048. https://doi.org/10.1093/nar/gky425 (2024) (July 2018).
24. Larralde, M. & Zeller, G. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* **39,** btad214. ISSN: 1367-4811. https://doi.org/10.1093/bioinformatics/btad214 (2024) (May 2023).

25. Johnson, M. C. *et al.* Core defense hotspots within Pseudomonas aeruginosa are a consistent and rich source of anti-phage defense systems. *Nucleic Acids Research* **51,** 4995–5005. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkad317 (2024) (June 2023).
26. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. eng. *Nucleic Acids Research* **53,** D672–D677. ISSN: 1362-4962 (Jan. 2025).
27. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. eng. *Nucleic Acids Research* **48,** D445–D453. ISSN: 1362-4962 (Jan. 2020).
28. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. en. *Nature Communications* **9.** Publisher: Nature Publishing Group, 2542. ISSN: 2041-1723. https://www.nature.com/articles/s41467-018-04964-5 (2024) (June 2018).
29. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. en. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008. ISSN: 1742-5468. https://dx.doi.org/10.1088/1742-5468/2008/10/P10008 (2025) (Oct. 2008).
30. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44,** D733–D745. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkv1189 (2024) (Jan. 2016).
31. Bokeh Development Team. *Bokeh: Python library for interactive visualization* 2018. https://bokeh.pydata.org/en/latest/.

Supplementary Figures

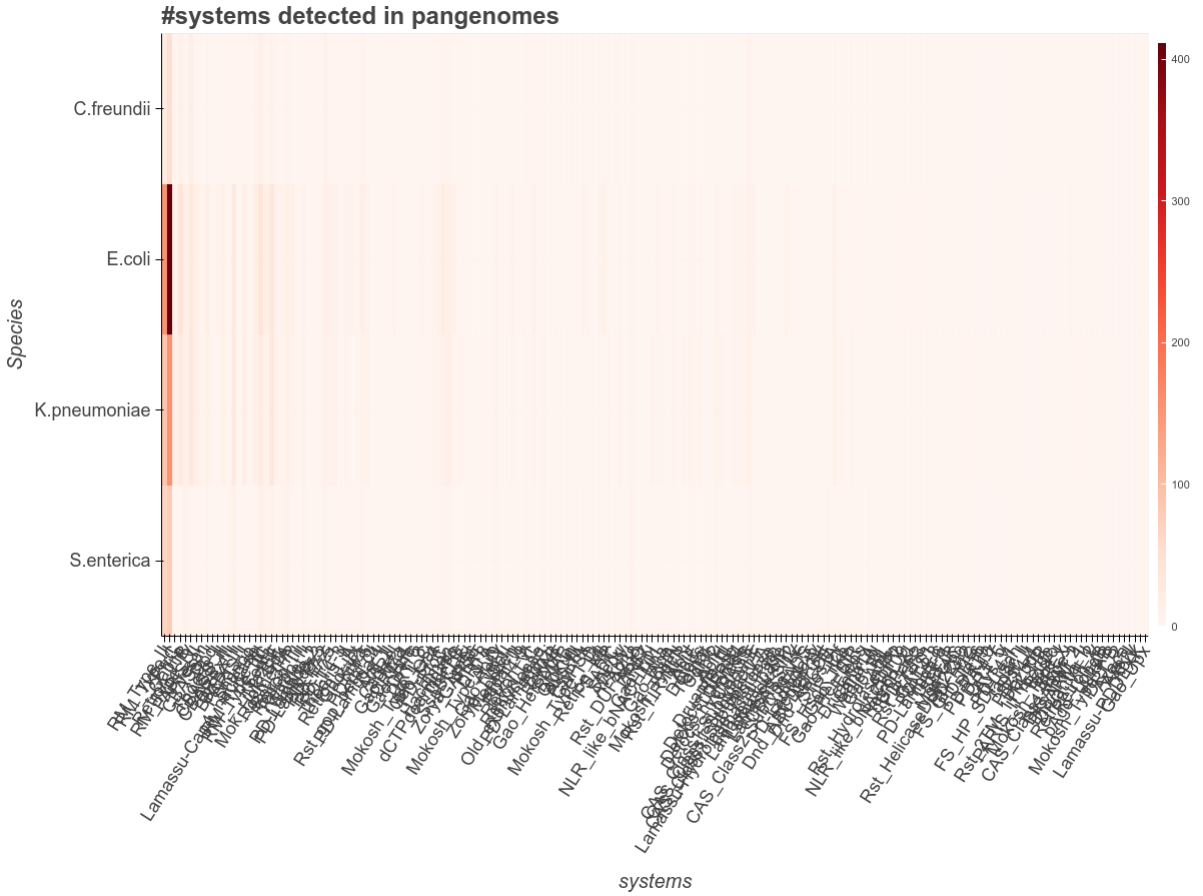

Fig. S1. Number of Systems Predicted for Each Model in Enterobacteriaceae Pangenomes. This figure is automatically generated by PANORAMA. PANORAMA employs the Bokeh package [31] to create interactive and dynamic visualizations, which can be directly accessed and manipulated in any standard web browser.

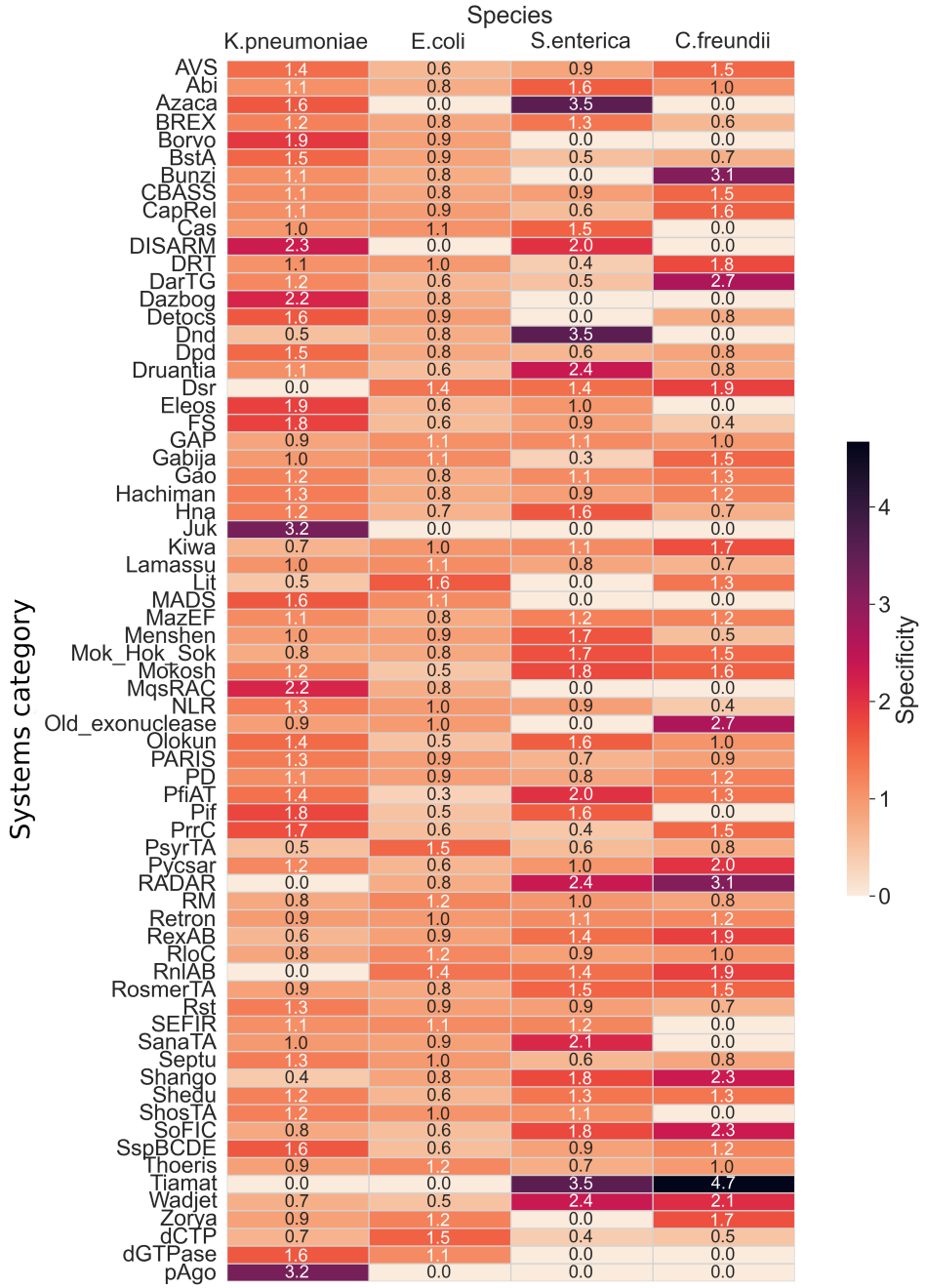

Fig. S2. Species-specificity evaluation across system categories in Enterobacteriaceae Pangenomes. The enrichment factors were computed using the method described in Equation 6, providing a quantitative measure of species-specific representation.

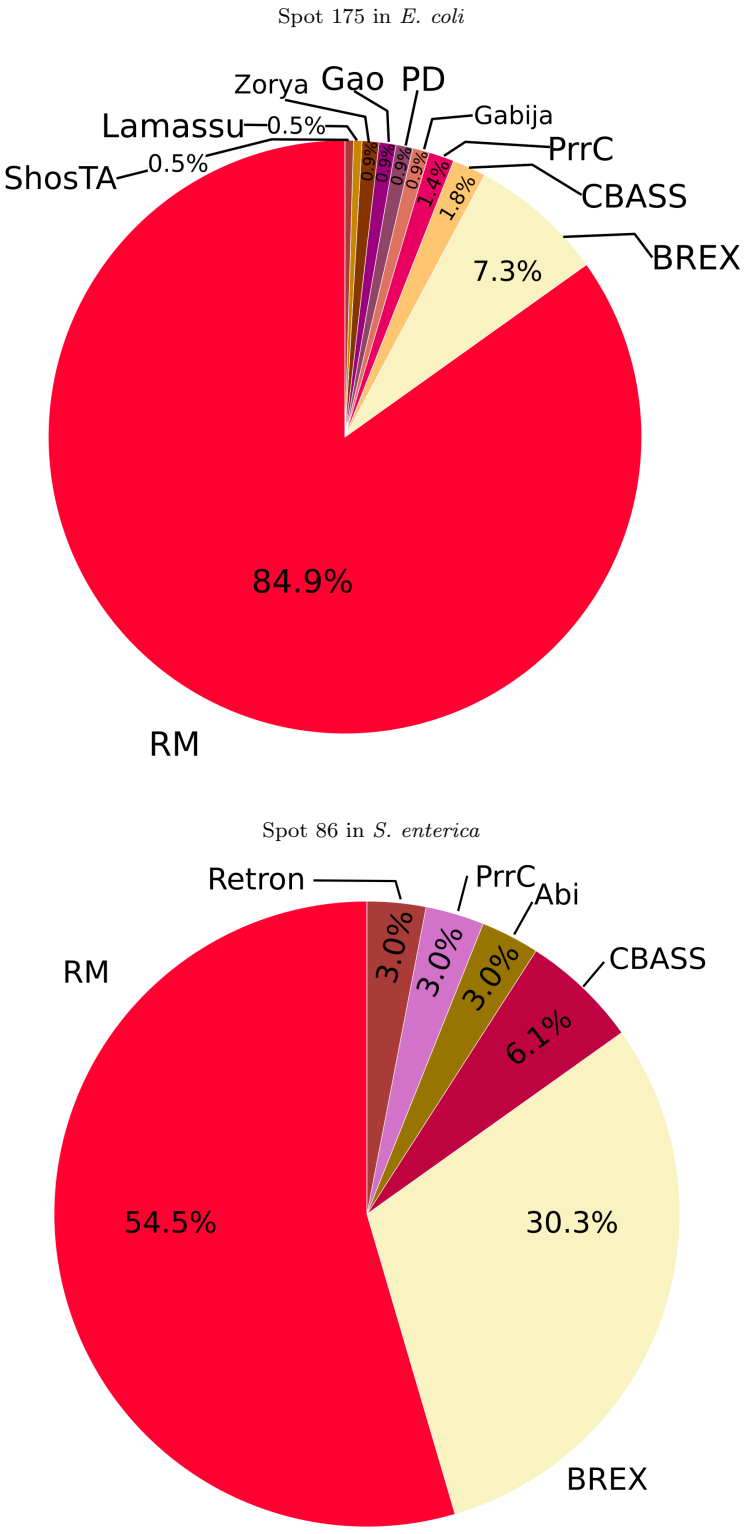

Fig. S3. Spot composition